MSE2122 - Nonlinear Optimization Lecture Notes IX

Fernando Dias (based on previous version by Fabricio Oliveira)

November 8, 2023

Abstract

In this lecture, we discuss the strategy of employing penalty functions to solve constrained optimisation problems. We also discuss the concept of penalised functions and demonstrate their asymptotical convergence properties. We consider a specific penalty method that has ties with Lagrangian duality, the augmented Lagrangian method of multipliers (ALMM). Moreover, we present a variant of ALMM, the alternating direction method of multipliers (ADMM), that allows for parallelisation of problems with suitable structure.

Contents

1 Penalty functions

The employment of penalty functions is a paradigm for solving constrained optimisation problems. The central idea of this paradigm is to convert the constrained optimisation problem into an unconstrained optimisation problem that is augmented with a *penalty function*, which penalises violations of the original constraints. The role of the penalty function is to allow steering the search towards feasible solutions in the search for optimal solutions.

Consider the problem (P) : min. $\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\}$. A *penalised version* of P is given by:

$$(P_\mu) : \text{min. } \{f(x) + \mu\alpha(x) : x \in X\},$$

where $\mu > 0$ is a *penalty term* and $\alpha(x) : \mathbb{R}^n \mapsto \mathbb{R}$ is a *penalty function* of the form:

$$\alpha(x) = \sum_{i=1}^{m} \phi(g_i(x)) + \sum_{i=1}^{l} \psi(h_i(x)). \tag{1}$$

For $\alpha(x)$ to be a suitable penalty function, one must observe that $\phi : \mathbb{R}^n \mapsto \mathbb{R}$ and $\psi : \mathbb{R}^n \mapsto \mathbb{R}$ are continuous and satisfy:

$$\phi(y) = 0 \text{ if } y \leq 0 \text{ and } \phi(y) > 0 \text{ if } y > 0$$
$$\psi(y) = 0 \text{ if } y = 0 \text{ and } \psi(y) > 0 \text{ if } y \neq 0.$$

Typical options are $\phi(y) = ([y]^+)^p$ with $p \in \mathbb{Z}_+$ and $\psi(y) = |y|^p$ with $p = 1$ or $p = 2$.

Figure 1 illustrates the solution of (P) : min. $\{x_1^2 + x_2^2 : x_1 + x_2 = 1, x \in \mathbb{R}^2\}$ using a penalty-based approach. Using $\alpha(x_1, x_2) = (x_1 + x_2 - 1)^2$, the penalised auxiliary problem P_μ becomes (P_μ) : min. $\{x_1^2 + x_2^2 + \mu(x_1 + x_2 - 1)^2 : x \in \mathbb{R}^2\}$. Since f_μ is convex and differentiable, necessary and sufficient optimality conditions $\nabla f_\mu(x) = 0$ imply:

$$x_1 + \mu(x_1 + x_2 - 1) = 0$$
$$x_2 + \mu(x_1 + x_2 - 1) = 0,$$

which gives $x_1 = x_2 = \frac{\mu}{2\mu+1}$.

One can notice that, as μ increases, the solution of the unconstrained penalised problem, represented by the level curves, becomes closer to the optimal of the original constrained problem P, represented by the dot on the hyperplane defined by $x_1 + x_2 = 1$.

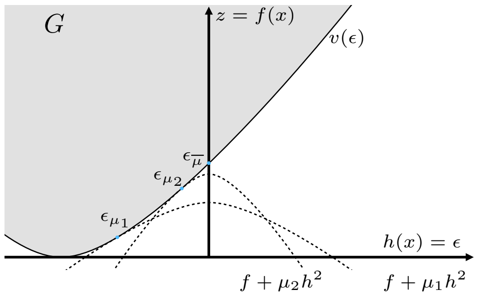

Figure 2: Geometric representation of penalised problems in the mapping $G = [h(x), f(x)]$

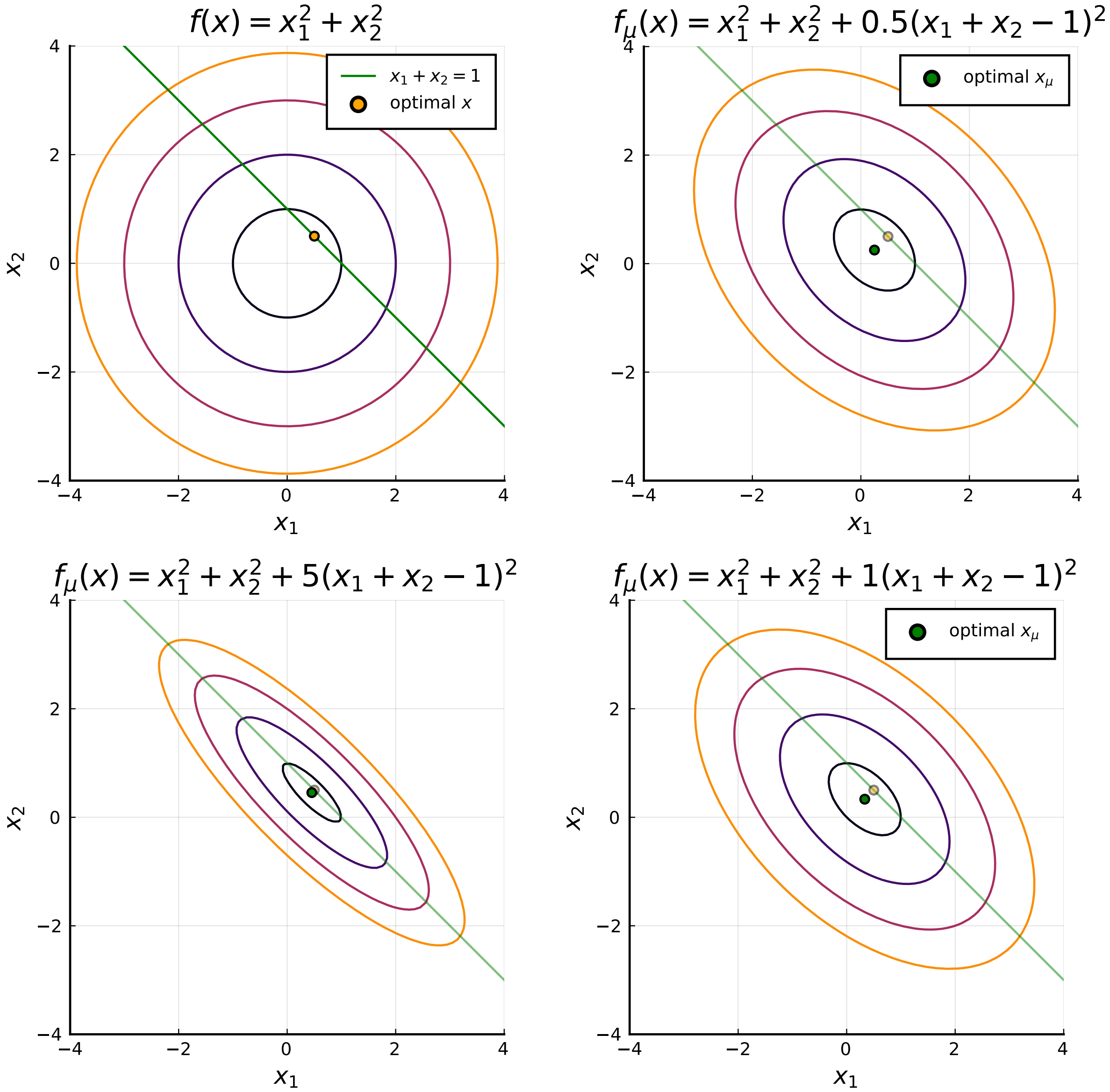

Figure 1: Solving the constrained problem P (top left) by gradually increasing the penalty term μ (0.5, 1, and 5, in clockwise order)

1.1 Geometric interpretation

A similar geometrical analysis to that performed with the Lagrangian duals can be employed for understanding how penalised problems can obtain optimal solutions. For that, let us the problem from the previous example $(P) : \text{min. } \{x_1^2 + x_2^2 : x_1 + x_2 = 1, x \in \mathbb{R}^2\}$. Let $G : \mathbb{R}^2 \to \mathbb{R}^2$ be a mapping $\{[h(x), f(x)] : x \in \mathbb{R}^2\}$, and let $v(\epsilon) = \text{min. } \{x_1^2 + x_2^2 : x_1 + x_2 - 1 = \epsilon, \ x \in \mathbb{R}^2\}$. The optimal solution is $x_1 = x_2 = \frac{1+\epsilon}{2}$ with $v(\epsilon) = \frac{(1+\epsilon)^2}{2}$.

Minimising $f(x) + \mu(h(x)^2)$ consists of moving the curve downwards until a single contact point ϵ_μ remains. One can notice that, as $\mu \to \infty$, $f + \mu h$ becomes sharper ($\mu_2 > \mu_1$), and ϵ_μ *converges* to the optimum $\epsilon_{\overline{\mu}}$. Figure 2 illustrates this behaviour.

The shape of the penalised problem curve is due to the following. First, notice that:

$$\begin{aligned}
&\min_x. \ \{f(x) + \mu \sum_{i=1}^{l} (h_i(x))^2\} \\
=&\min_{x,\epsilon}. \ \{f(x) + \mu||\epsilon||^2 : h_i(x) = \epsilon, i = 1, \dots, l\} \\
=&\min_{\epsilon}. \ \{\mu||\epsilon||^2 + \min_x. \ \{f(x) : h_i(x) = \epsilon, i = 1, \dots, l\}\} \\
=&\min_{\epsilon}. \ \{\mu||\epsilon||^2 + v(\epsilon)\}.
\end{aligned}$$

Consider $l = 1$, and let $x_\mu = \mathbf{argmin}_x\{f(x) + \mu \sum_{i=1}^{l}(h_i(x))^2\}$ with $h(x_\mu) = \epsilon_\mu$, implying that $\epsilon_\mu = \mathbf{argmin}_\epsilon\{\mu||\epsilon||^2 + v(\epsilon)\}$. Then, the following holds:

1. $f(x_\mu) + \mu(h(x_\mu))^2 = \mu\epsilon_\mu^2 + v(\epsilon_\mu) \Rightarrow f(x_\mu) = v(\epsilon_\mu)$, since $h(x_\mu) = \epsilon_\mu$;
2. and $v'(\epsilon_\mu) = \frac{\partial}{\partial \epsilon}(f(x_\mu) + \mu(h(x_\mu))^2 - \mu\epsilon_\mu^2) = -2\mu\epsilon_\mu$.

Therefore, $(h(x_\mu), f(x_\mu)) = (\epsilon_\mu, v(\epsilon_\mu))$. Denoting $f(x_\mu) + \mu h(x_\mu)^2 = k_\mu$, we see the parabolic function $f = k_\mu - \mu\epsilon^2$ matching $v(\epsilon_\mu)$ for $\epsilon = \epsilon_\mu$ and has the slope $-2\mu\epsilon$, matching that of $v(\epsilon)$ at that point.

1.2 Penalty function methods

The convergent behaviour of the penalised problem as the penalty term μ increases inspires the development of a simple yet powerful method for optimising constrained optimisation problems.

That is, consider the problem P defined as:

$$\begin{aligned}
(P) : \min. \ & f(x) \\
& g_i(x) \le 0, i = 1, \dots, m, \\
& h_i(x) = 0, i = 1, \dots, l, \\
& x \in X.
\end{aligned}$$

We seek to solve P by solving $\sup_\mu\{\theta(\mu)\}$ for $\mu > 0$, where:

$$\theta(\mu) = \min. \ \{f(x) + \mu\alpha(x) : x \in X\}$$

and $\alpha(x)$ is a penalty function as defined in (1). For that to be possible, we need first to state a convergence result guaranteeing that:

$$\min. \ \{f(x) : g(x) \le 0, h(x) = 0, x \in X\} = \sup_{\mu \ge 0} \theta(\mu) = \lim_{\mu \to \infty} \theta(\mu).$$

In practice, that would mean that μ_k can be increased at each iteration k until a suitable tolerance is achieved. Theorem 1.1 states the convergence of penalty based methods.

Theorem 1.1. Convergence of penalty-based methods

Consider the problem P, where f, g_i for $i = 1, \dots, m$, and h_i for $i = 1, \dots, l$ are continuous, and $X \subset \mathbb{R}^n$ a compact set. Suppose that, for each μ, there exists $x_\mu = \mathbf{argmin}\{f(x) + \alpha(x) : x \in X\}$, where α is a suitable penalty function and $\{x_\mu\}$ is contained within X. Then:

$$\min_x\{f(x) : g(x) \le 0, h(x) = 0, x \in X\} = \sup_{\mu \ge 0}\{\theta(\mu)\} = \lim_{\mu \to \infty} \theta(\mu),$$

where $\theta(\mu) = \min_x\{f(x) + \mu\alpha(x) : x \in X\} = f(x_\mu) + \mu\alpha(x_\mu)$. Also, the limit of any convergent subsequence of $\{x_\mu\}$ is optimal to the original problem and $\mu\alpha(x_\mu) \to 0$ as $\mu \to \infty$.

Proof. We first show that $\theta(\mu)$ are nondecreasing function of μ. Let $0 < \lambda < \mu$. From the definition of $\theta(\mu)$, we have that:

$$f(x_\mu) + \lambda\alpha(x_\mu) \geq f(x_\lambda) + \lambda\alpha(x_\lambda) \tag{2}$$

Adding and subtracting $\mu\alpha(x_\mu)$ in the left side of (2), we conclude that $\theta(\mu) \geq \theta(\lambda)$. Now, for $x \in X$ with $g(x) \leq 0$ and $h(x) = 0$, notice that $\alpha(x) = 0$. This implies that:

$$f(x) = f(x) + \mu\alpha(x) \geq \inf_x\{f(x) + \mu\alpha(x) : x \in X\} = \theta(\mu) \tag{3}$$

and, therefore, $\theta(\mu)$ is bounded above, and thus $\sup_{\mu\geq 0}\theta(\mu) = \lim_{\mu\to\infty}\theta(\mu)$. For that to be the case, we must have that $\mu\alpha(x_\mu) \to 0$ as $\mu \to \infty$. Moreover, we notice from (3) that:

$$\min_x\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\} \geq \lim_{\mu\to\infty}\theta(\mu). \tag{4}$$

On the other hand, take any convergent subsequence $\{x_{\mu_k}\}$ of $\{x_\mu\}_{\mu\to\infty}$ with limit $\overline{x}$. Then:

$$\sup_{\mu\geq 0}\theta(\mu) \geq \theta(\mu_k) = f(x_{\mu_k}) + \mu\alpha(x_{\mu_k}) \geq f(x_{\mu_k}).$$

Since $x_{\mu_k} \to \overline{x}$ as $\mu \to \infty$ and f is continuous, this implies that $\sup_{\mu\geq 0}\theta(\mu) \geq f(\overline{x})$. Combined with (4), we have that $f(\overline{x}) = \sup_{\mu\geq 0}\{\theta(\mu)\}$ and thus the result follows.

😎

The proof starts by demonstrating the nonincreasing behaviour of penalty functions and nondecreasing behaviour of $\theta(\mu)$ to allow for convergence. By noticing that:

$$f(x_\mu) + \lambda\alpha(x_\mu) + \mu\alpha(x_\mu) - \mu\alpha(x_\mu) = \theta(\mu) + (\lambda - \mu)\alpha(x_\mu) \geq f(x_\lambda) + \lambda\alpha(x_\lambda) = \theta(\lambda)$$

and that $\lambda - \mu < 0$, we can infer that $\theta(\mu) \geq \theta(\lambda)$. It is also interesting to notice how the objective function $f(x)$ and infeasibility $\alpha(x)$ behave as we increase the penalty coefficient μ. For that, notice that using the same trick in the proof for two distinct values $0 < \lambda < \mu$, we have:

1. $f(x_\mu) + \lambda\alpha(x_\mu) \geq f(x_\lambda) + \lambda\alpha(x_\lambda)$
2. $f(x_\lambda) + \mu\alpha(x_\lambda) \geq f(x_\mu) + \mu\alpha(x_\mu)$.

Notice that in 1, we use the fact that $x_\lambda = \mathbf{argmin}_x\theta(\lambda) = \mathbf{argmin}_x\{f(x) + \lambda\alpha(x)\}$ and therefore, must be less or equal then $f(x_\mu) + \lambda\alpha(x_\mu)$ for an arbitrary $x_\mu \in X$. The same logic is employed in 2, but reversed in λ and μ. Adding 1 and 2, we obtain $(\mu - \lambda)(\alpha(x_\lambda) - \alpha(x_\mu)) \geq 0$ and conclude that $\alpha(x_\mu) \leq \alpha(x_\lambda)$ for $\mu > \lambda$, i.e., that $\alpha(x)$ is nonincreasing in μ.

Moreover, from the first inequality, we have that $f(x_\mu) \geq f(x_\lambda)$. Notice how this goes in line with what one would expect from the method: as we increase the penalty coefficient μ, the optimal infeasibility, measured by $\alpha(x_\mu)$ decreases, while the objective function value $f(x_\mu)$ worsens at it is slowly "forced" to be closer to the original feasible region.

Note that the assumption of compactness plays a central role in this proof, such that $\theta(\mu)$ can be evaluated for any μ as $\mu \to \infty$. Though this is a strong assumption, it tends to not be so restrictive in practical cases, since variables typically lie within finite lower and upper bounds. Finally, notice that $\alpha(\overline{x}) = 0$ implies that $\overline{x}$ is feasible for g_i for $i = 1, \dots, m$, and h_i for $i = 1, \dots, l$, and thus optimal for P. This is stated in the following corollary.

Corollary 1.2. If $\alpha(x_\mu) = 0$ for some μ, then x_μ is optimal for P.

Proof. If $\alpha(x_\mu) = 0$, then x_μ is feasible. Moreover, x_μ is optimal, since

$$\begin{aligned}\theta(\mu) &= f(x_\mu) + \mu\alpha(x_\mu)\\ &= f(x_\mu) \leq \inf\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\}.\end{aligned}$$

😎

A technical detail of the proof of Theorem 1.1 is that the convergence of such approach is asymptotically, i.e., by making μ arbitrarily large, x_μ can be made arbitrarily close to the true optimal $\overline{x}$ and $\theta(\mu)$ can be made arbitrarily close to the optimal value $f(\overline{x})$. In practice, this strategy tends to be prone to computational instability.

The computational instability arises from the influence that the penalty term exerts in some of the eigenvalues of the Hessian of the penalised problem. Let $H_\mu(x_\mu)$ be the Hessian of the penalised function at x_μ. Recall that conditioning is measured by $\kappa = \frac{\max_{i=1,\dots,n}\lambda_i}{\min_{i=1,\dots,n}\lambda_i}$, where $\{\lambda_i\}_{i=1,\dots,n}$ are the eigenvalues of $H_\mu(x_\mu)$. Since the influence is only on some of the eigenvalues, this affects the conditioning of the problem and might lead to numerical instabilities. An indication of that can be seen in Figure 1, where one can notice the elongated profile of the function as the penalty term μ increases.

Consider the following example. Let the penalised function $f_\mu(x) = x_1^2 + x_2^2 + \mu(x_1 + x_2 - 1)^2$.

The Hessian of $f_\mu(x)$ is:

$$\nabla^2 f_\mu(x) = \begin{bmatrix} 2(1+\mu) & 2\mu \\ 2\mu & 2(1+\mu) \end{bmatrix}.$$

Solving $\mathbf{det}(\nabla^2 f_\mu(x) - \lambda I) = 0$, we obtain $\lambda_1 = 2$, $\lambda_2 = 2(1+2\mu)$, with eigenvectors $(1,-1)$ and $(1,1)$, which gives $\kappa = (1+2\mu)$. This illustrates that the eigenvalues, and consequently the conditioning number, is proportional to the penalty term.

2 Augmented Lagrangian method of multipliers

For simplicity, consider the (primal) problem P as (P) : min. $\{f(x) : h_i(x) = 0,\ i = 1, \dots, l\}$. The augmented Lagrangian method of multipliers arises from the idea of seeking for a penalty term that would allow for exact convergence for a finite penalty.

Considering the geometrical interpretation in Figure 2, one might notice that a horizontal shift in the penalty curve would allow for the extreme point of the curve to match the optimum on the z ordinate.

Therefore, we consider a modified penalised problem of the form:

$$f_\mu(x) = f(x) + \mu \sum_{i=1}^{l} (h_i(x) - \theta_i)^2$$

where θ_i is the shift term. One can notice that:

$$\begin{aligned} f_\mu(x) &= f(x) + \mu \sum_{i=1}^{l} (h_i(x) - \theta_i)^2 \\ &= f(x) + \mu \sum_{i=1}^{l} h_i(x)^2 - \sum_{i=1}^{l} 2\mu\theta_i h_i(x) + \mu \sum_{i=1}^{l} \theta_i^2 \\ &= f(x) + \sum_{i=1}^{l} v_i h_i(x) + \mu \sum_{i=1}^{l} h_i(x)^2, \end{aligned}$$

with $v_i = -2\mu\theta_i$. The last term is a constant and can be dropped.

The term *augmented Lagrangian* refers to the fact that $f_\mu(x)$ is equivalent to the Lagrangian function of problem P, augmented by the penalty term.

This allows for noticing important properties associated with the augmented Lagrangian $f_\mu(x)$. For example, assume that $(\overline{x}, \overline{v})$ is a KKT solution to P. Then the optimality condition:

$$\nabla_x f_\mu(x) = \nabla f(x) + \sum_{i=1}^{l} \overline{v}_i \nabla h_i(x) + 2\mu \sum_{i=1}^{l} h_i(x) \nabla h_i(x) = 0,$$

implies that the optimal solution $\overline{x}$ can be recovered using a finite penalty term μ, unlike with the previous penalty-based method. The existence of finite penalty terms $\mu > 0$ that can recover optimality has an interesting geometrical interpretation, in light of what was previously discussed. Consider the same setting from Figure 2, but now we consider curves of the form $f + \overline{v}h + \mu h^2 = k$. This is illustrated in Figure 3.

Optimising the augmented Lagrangian function amounts to finding the curve $f + \overline{v}h + \mu h^2 = k$ in which $v(\epsilon) = k$. The expression for k can be conveniently rewritten as $f = -\mu \left[h + (\overline{v}/2\mu)\right]^2 + \left[k + (\overline{v}^2/4\mu)\right]$, exposing that f is a parabola shifted by $h = -\overline{v}/2\mu$.

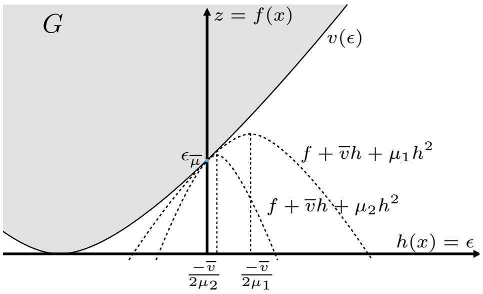
G
$z = f(x)$
$v(\epsilon)$
$\epsilon_{\overline{\mu}}$
$f + \overline{v}h + \mu_1 h^2$
$f + \overline{v}h + \mu_2 h^2$
$h(x) = \epsilon$
$\frac{-\overline{v}}{2\mu_2}$
$\frac{-\overline{v}}{2\mu_1}$

Figure 3: Geometric representation of augmented Lagrangians in the mapping $G = [h(x), f(x)]$

2.1 Augmented Lagrangian method of multipliers

We can employ an unconstrained optimisation method to solve the augmented Lagrangian function:

$$L_\mu(x, v) = f(x) + \sum_{i=1}^{l} v_i h_i(x) + \mu \sum_{i=1}^{l} h_i(x)^2,$$

which amount to rely on strong duality and search for KKT points (or primal-dual pairs) $(\overline{x}, \overline{v})$ by iteratively operating in the primal (x) and dual (v) spaces. In particular, the strategy consists of:

1. *Primal space:* optimise $L_\mu(x, v^k)$ using an unconstrained optimisation method;

2. *Dual space:* perform a dual variable update step retaining the optimality condition $\nabla_x L(x^{k+1}, v^k) = \nabla_x L(x^{k+1}, v^{k+1}) = 0$.

This strategy is akin to applying the subgradient method to solving the augmented Lagrangian dual. The update step for the dual variable is given by:

$$\overline{v}^{k+1} = \overline{v}^k + 2\mu h(\overline{x}^{k+1}).$$

The motivation for the dual step update stems from the following observation:

1. $h(\overline{x}^k)$ is a subgradient of $L_\mu(x, v)$ at $\overline{x}^k$ for any v.

2. The step size is devised such that the optimality condition of the Lagrangian function is retained, i.e., $\nabla_x L(\overline{x}^k, \overline{v}^{k+1}) = 0$.

Part 2 refers to the following:

$$\begin{aligned}\nabla_x L(\overline{x}^k, \overline{v}^{k+1}) &= \nabla f(\overline{x}^k) + \sum_{i=1}^{l} \overline{v}_i^{k+1} \nabla h_i(\overline{x}^k) = 0 \\ &= \nabla f(\overline{x}^k) + \sum_{i=1}^{l} (\overline{v}_i^k + 2\mu h_i(\overline{x}^k)) \nabla h_i(\overline{x}^k) = 0 \\ &= \nabla f(\overline{x}^k) + \sum_{i=1}^{l} \overline{v}_i^k \nabla h_i(\overline{x}^k) + \sum_{i=1}^{l} 2\mu h_i(\overline{x}^k) \nabla h_i(\overline{x}^k) = \nabla_x L_\mu(\overline{x}^k, \overline{v}^k) = 0.\end{aligned}$$

That is, by employing $\overline{v}^{k+1} = \overline{v}^k + 2\mu h(\overline{x}^{k+1})$ one can retain optimality in the dual variable space for the Lagrangian function from the optimality conditions of the penalised functions, which is a condition for $\overline{x}$ to be a KKT point.

Algorithm 1 summarises the augmented Lagrangian method of multipliers (ALMM).

Algorithm 1 Augmented Lagrangian method of multipliers (ALMM)

```
initialise. tolerance $\epsilon > 0$, initial dual solution $v^0$, iteration count $k = 0$
while $|h(\overline{x}^k)| > \epsilon$ do
    $\overline{x}^{k+1} = \mathbf{argmin} L_\mu(x, \overline{v}^k)$
    $\overline{v}^{k+1} = \overline{v}^k + 2\mu h(\overline{x}^{k+1})$
    $k = k + 1$
end while
return $x^k$.
```

The method can be specialised such that μ is individualised for each constraint and updated proportionally to the observed infeasibility $h_i(x)$. Such a procedure is still guaranteed to converge, as the requirement in Theorem 1.1 that $\mu \to \infty$ is still trivially satisfied.

One important point about the augmented Lagrangian method of multipliers is that linear convergence is to be expected, due to the gradient-like step taken to find optimal dual variables. This is often the case with traditional Lagrangian duality based approaches.

2.2 Alternating direction method of multipliers - ADMM

ADMM is a distributed version of the augmented Lagrangian method of multipliers, and is more suited to large problems with a decomposable structure.

Consider the problem P to be of the following form:

$$\begin{aligned} (P) : \text{min. } & f(x) + g(y) \\ \text{subject to: } & Ax + By = c. \end{aligned}$$

We would like to be able to solve the problem separately for x and y, which could, in principle be achieved using ALMM. However, the consideration of the penalty term prevents the problem from being completely separable. To see that, let

$$\phi(x, y, v) = f(x) + g(y) + v^\top(c - Ax - By) + \mu(c - Ax - By)^2$$

be the augmented Lagrangian function. One can notice that the penalty term $\mu(c - Ax - By)^2$ prevents the separation of the problem in terms of the x and y variables. However, separability can be recovered is one employs a coordinate descent approach in which three blocks are considered: x, y, and v. The ADMM is summarised in Algorithm 2.

Algorithm 2 ADMM

```
1: initialise. tolerance $\epsilon > 0$, initial dual and primal solutions $v^0$ and $y^0$, $k = 0$
2: while $|c - A\overline{x}^k - B\overline{y}^k|$ and $||y^{k+1} - y^k|| > \epsilon$ do
3:     $\overline{x}^{k+1} = \mathbf{argmin} \phi_\mu(x, \overline{y}^k, \overline{v}^k)$
4:     $\overline{y}^{k+1} = \mathbf{argmin} \phi_\mu(\overline{x}^{k+1}, y, \overline{v}^k)$
5:     $\overline{v}^{k+1} = \overline{v}^k + 2\mu(c - A\overline{x}^{k+1} - B\overline{y}^{k+1})$
6:     $k = k + 1$
7: end while
8: return $(x^k, y^k)$.
```

One important feature regarding ADMM is that the coordinate descent steps are taken in a cyclic order, not requiring more than one (x, y) update step. Variants consider more than one of these steps, but no clear benefit in practice has been observed. Moreover, μ can be updated according to the amount of infeasibility observed at iteration k, but no generally good update rule is known.

ADMM is particularly relevant as a method for (un)constrained problems in which it might expose a structure that can be exploited, such as having in some of the optimisation problems (in Lines 3 and 4 in Algorithm 2) that might have solutions in closed forms.